GRAF Software Documentation

Version 2.3

Yumi (Jimmy) Jin, PhD
jinyu@ncbi.nlm.nih.gov

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

4/17/2018

OVERVIEW

GRAF (**g**enetic **r**elationship **a**nd **f**ingerprinting) is a package to do some useful analyses and visualization of dbGaP data (Jin et al, 2017). The package GRAF (upper case) includes the program graf (lower case), which is a C++ program that quickly finds the closely related subjects using SNP genotype data. The main program, graf, can be used to validate the relationships of subjects listed in the pedigree file by analyzing the subject-sample mapping (SSM) file and the genotype data for the samples.

The main program, graf, extracts genotypes of a fixed set of 10,000 dbGaP fingerprinting SNPs from the input genotype datasets (i.e., PLINK binary .bed files) and uses these genotypes to check each pair of samples to determine 1) if the two samples are collected from one subject (or from two identical twins) or two different subjects; 2) whether or not the two subjects are closely related. graf can read genotypes from datasets obtained with different genotyping methods and find duplicate samples and related subjects in these datasets.

graf compares the genotypes of all the 10,000 fingerprinting SNPs (distributed as the file FP_SNPs.txt) and calculates the **a**ll **g**enotype **m**ismatch **r**ate (AGMR) and the **h**omozygous **g**enotype **m**ismatch **r**ate (HGMR) for each pair of samples. AGMR is the percentage of SNPs on which the two genotypes are not identical, while HGMR is the genotype mismatch rate when only the SNPs with homozygous calls for both samples are considered.

More information on how the computations are done is provided in the section below entitled METHODS.

UPDATES:

04/06/2017 version 1.1:

A multithreaded C++ program *graf_dups* is added to the package to quickly find duplicate samples and monozygotic twins. Please run *graf_dups -help* for description and usage.

10/10/2017 version 2.0:

The GRAF-pop feature is added into the package, which includes a new *-pop* option of the graf program and a Perl script *PlotPopulations.pl.* See GRAF-popDocumentation_YYYYMMDD.docx for descriptions of the new feature, which is not included in this document, except that the subsection below entitled "Reporting Questions and Problems" is generally applicable. YYYYMMDD represents a date.

12/12/2017 version 2.1:

GRAF-pop is improved to tolerate non-random missing genotypes.

02/26/2018 version 2.2:

The C++ program *graf* (including the GRAF-pop feature) is made multithreaded.

An option *-areas* is added to script *PlotPopulations.pl* for user to see the expected areas on the population scatter plot that include 95% of the dbGaP subjects with sufficient large number of fingerprint SNPs with genotypes.

An error in the equation to calculate genetic distance in GRAF-popDocumentation is fixed.

04/17/2018 version 2.3:

Statistical scores are renamed as GD1, GD2, GD3, GD4 (See the GRAF-pop documentation).

Algorithms to assign populations based on statistical scores are modified.

Functions are added *PlotPopulations.pl* to (1) plot graph of GD1 vs. GD4; (2) plot cutoff lines that are used to classify subjects into pre-defined populations.

INPUT FILES AND OTHER FILES IN THE GRAF DISTRIBUTION

Within the file GrafPkg.tar.gz, the GRAF package includes four files intended to be executable: the C++ programs *graf*, *graf_dups*, and the perl scripts *PlotGraf.pl, PlotPopulatioin.pl* that are visible as separate files after the user executes the commands

```
gunzip GrafPkg.tar.gz
tar xvf GrafPkg.tar
```

The four executable programs are *graf, graf_dups, PlotGraf.pl* and *PlotPopulations.pl*.

graf compares all pairs of subjects and finds and reports the closely related pairs, while *PlotGraf.pl* takes the file generated by *graf* and plots graphs to show the distributions of HGMR and AGMR values. Note that *PlotGraf.pl* and *PlotPopulations.pl* require that GD Graphics Library (http://search.cpan.org/~lds/GD-1.38/GD.pm) be installed.

In most usages, *graf* expects as input one or more genotype datasets in PLINK format, i.e., .bed, .bim and .fam files that share a prefix in their names. However, since multiple samples can be collected from one subject and the subject-sample mapping information is not stored in datasets in PLINK, *graf* reads subject-sample mapping and pedigree information from the dbGaP SSM file and pedigree file. The IDs (second column, no column header) in the PLINK .fam file are read as sample IDs by *graf*. If subject IDs are the same as sample IDs, then no SSM file is necessary, and *graf* will read the pedigree information from the PLINK .fam file.

However, if any of the sample IDs are different from their corresponding subject IDs, then an SSM file should be passed to *graf*. The SSM file should be a tab-delimited plain text file with a sample column and a subject column (with column headers, see dbGaP submission guide). When an SSM file is provided, a pedigree file (see dbGaP submission guide) should also be provided to pass the pedigree information to *graf*. The pedigree file should be a tab-delimited plain text file with at least the following 5 columns (with a column header row):

Column 1: FamilyID
Column 2: SubjectID
Column 3: FatherID
Column 4: MotherID
Column 5: Sex (1 = male; 2 = female; 0 or NULL = unknown)

SubjectID, FatherID and MotherID are IDs of subjects, not samples.

The SSM format is a two-column tab delimited text file that establishes a mapping from Sample IDs to Subject IDs. The columns should have the headers Subject_ID and Sample_ID, respectively. An example SSM format file is included in the GRAF distribution with the name *affy_hapmap_ssm.txt*.

If there are identical twins in the datasets, the twin information should be entered to the optional 6th column "TwinID", where the same twin ID (can be an integer or a string) is used to indicate that subjects are identical twins. For example, if three subjects A, B, C are identical triplets, a unique subject ID, e.g., the integer 18, can be created for them and entered into the TwinID column for subjects A, B, C.

The sample genotypes can also be stored in datasets with GRAF format. *graf* uses a single .fpg file to store the sample genotypes. A .fpg file is a plain text file with three columns: the first column is the dataset ID (integer) column, the second one is the sample ID column, and the third column stores sample genotypes in strings of hexadecimal numbers. Each hexadecimal number represents genotypes of two fingerprinting SNPs. The first hexadecimal number stores genotypes of the first two fingerprinting SNPs, the second number keeps genotypes of fingerprinting SNPs #3 and #4, and so on. If the hexadecimal number is converted to a binary number, then the first two bits keep the genotype of the first SNP and the last two bits are for the second SNP, with the following code meanings:

00: 0 reference alleles
01: 1 reference allele
10: 2 reference alleles
11: missing genotype

The. fpg file can be generated using the -geno option of the *graf* program and reused as input to the program in a subsequent run.

Included in the distribution are two sample datasets for which the file names have prefixes *affy_hapmap* and *perlegen_hapmap*. Both sets of sample files come in byte-encode PLINK format meaning that there are three files with suffixes {fam,bim,bed}.

RUNNING GRAF

graf is a command line executable that can be run under GNU/LINUX 64 bit systems. Brief instructions are given when the program is executed without parameters:

```
graf

Usage: graf [options]
    -plink  PLINK set root:  File root of PLINK .bed, .bim and .fam
                             files
    -geno   fpg file:        Specify GRAF .fpg file
    -exfp   PLINK set list:  Extract fingerprinting genotypes from a
                             list of PLINK sets (file roots) separated
                             by commas
    -pop    output file:     Check subject populations and save results
                             to the output file
    -out    output file:     Output file to save the results
    -appd   DS No.:          Append extracted fingerprinting genotypes
                             to the output file.  The integer is
                             dataset No. of the first PLINK set
    -ssrs   SS-RS mapping:   Specify SS# to RS# mapping file (Two
                             columns: SS# and RS# without column
                             headers)
    -ped    pedigree file:   Specify pedigree file of subject IDs (with
                             column headers)
    -ssm    SSM file:        Specify dbGaP subject-sample mapping file
    -maxhm  max HGMR value:  Specify maximum HGMR values for a pair of
                             subjects to be reported by GRAF
    -xpmr   type:            Specify how expected HGMR and AGMR values
                             are calculated for each type of relationship
                             (default 1)
                             1: Use input dataset to calculate the
                                expected HGMR and AGMR values
                             2: Use average HGMR and AGMR values in dbGaP
                                database for the expected values
    -type   relation_type:   Specify relation type.  Acceptable values
                             are 1, 2, 3, or 4 (default 3)
                             1: Find all duplicates and PO pairs
                             2: Find all duplicates, PO and FS pairs
                             3: Find all duplicates, PO, FS and second
                                degree relatives
                             4: Compare all the 10,000 SNPs to find all
                                the related subjects

NOTE:
    1. Exactly one of the following two options should be selected:
       -plink or –geno.
    2. When option -exfp is selected, -out must also be selected and
       output file should have .fpg extension.
    3. When multiple PLINK sets are used, each dataset will be assigned
       an integer dataset ID starting with 1.
    4. The above PLINK set starting index can be specified using option
       -appd. When -appd is selected, the out file should be an existing
       GRAF .fpg file.
   5. Multiple datasets can be combined into a single geno file using
       the –exfp and –appd options.
   6. When multiple datasets are used, both inter- and intra- dataset
       pairwise comparisons are done. There is currently no way to
       restrict to one category of comparisons or the other.
```

Below are more detailed descriptions (with examples) of these options.

-plink
Allows the user to specify the name of the genotype dataset in PLINK .bed, .bim, .fam format. The parameter should be the file root of the plink set. In this example, *graf* will try to find the following three files: affy_hapmap.bed, aff_hapmap.bim and affy_hapmap.fam.

Example: `graf -plink affy_hapmap`

-exfp
Extracts fingerprinting genotypes from multiple PLINK sets and saves the results to the file name specified by -out option. The datasets will be given integer dataset IDs starting from 1. The output file name should be new.

Example: `graf -exfp affy_hapmap,perlegen_hapmap -out comb_hapmap.fpg`

-pop

Checks subject populations using the fingerprinting genotypes and saves the results to the output file.

Example: `graf -plink G1000FpGeno -pop G1000_sbj_scores.txt`

See the related documentation in GRAF-popDocumntation_YYYYMMDD.docx, where YYYYMMDD represents a date.

-exfp –appd

Extracts fingerprinting genotypes from a PLINK set and appends the results to an existing output file, with dataset ID specified by –appd option.

Example:

Step 1: `graf -exfp affy_hapmap -out comb_hapmap2.fpg`

Step 2: `graf -exfp perlegen_hapmap -out comb_hapmap2.fpg –appd 2`

-geno

Allows the user to specify the name of the genotype dataset in GRAF format.

Example: `graf -geno comb_hapmap.fpg`

-ssm

Allows the user to specify the name of the subject-sample mapping file in dbGaP format. When sample IDs are different from subject IDs, a subject-sample mapping file is required. The subject-sample mapping file should list all the sample IDs in the PLINK .fam file and their corresponding subject IDs.

Example: `graf -plink affy_hapmap -ssm affy_hapmap_ssm.txt`

-ped

Allows the user to specify the pedigree file in dbGaP format. When pedigree file is specified with -ped option, *graf* will ignore the pedigree information in the PLINK .fam file and read the information from the pedigree file. The IDs in the pedigree file should be subject IDs. This option can take only one dataset at a time.

Example:
```
graf -plink affy_hapmap -ssm affy_hapmap_ssm.txt
-ped affy_hapmap_fake_pedigree.txt
```

-out

Allows the user to specify the name of the output file for saving the related pairs of samples detected by *graf*. If the output file is not specified, the output will be saved to a default file graf_rel_yyyymmdd_hhmm.txt, where yyyymmdd_hhmm is the current local time in this format.

Example: `graf -plink affy_hapmap –out aff_hapmap_rels.txt`

-maxhm

Sets the maximum HGMR value for related pairs outputted by *graf*. Subject pairs with HGMR greater than this value will be treated by *graf* as unrelated and will not be saved to the output file. The default maximum HGMR is 20.

Example: `graf -plink affy_hapmap –out aff_hapmap_rels.txt –maxhm 15`

-xpmr

Allows the user to specify how the expected HGMR and AGMR values are calculated. For each pair of subjects, GRAF estimates the allele frequency distribution of the fingerprinting SNPs of the population where the subjects are sampled, and then uses these allele frequencies to calculate the expected HGMR and AGMR values. Assuming all of the subjects in the input file(s) are sampled from the same population, GRAF uses the allele frequencies of all subjects in the input datasets to estimate the allele frequencies in the population. In cases when the sample size is small (fewer than 100 subjects) in the input datasets, GRAF uses the allele frequencies of all the subjects in dbGaP Fingerprint Collection to estimate the population allele frequencies. The user can use -xpmr option (1 or 2) to let GRAF choose one of the above two options to estimate the population allele frequencies. When the selection –xmpr 1 is combined with choices of –geno or –exfp that combine multiple datasets, then the allele frequencies are combined as a weighted average of all the participating datasets and the same weighted average is used for all pairwise comparisons.

Example: `graf -plink affy_hapmap –xpmr 2`

-type

Usage of *graf* involves a tradeoff between running time and prediction accuracy. To obtain high sensitivity, the program needs to check more SNPs, at the expense of a longer running time. The –type option allows the user to specify the relative type for which *graf* should try to find all the pairs. The type should be an integer from 1 to 4, with the code meanings shown in the above short description. The greater the type value is, the more SNPs *graf* will check, and hence the more related samples it will find and the more time it will spend. The default type value is 3.

Example: `graf -plink affy_hapmap –type 2`

-ssrs

When the marker IDs in the PLINK .bim file are SS IDs, the user can use -ssrs option to specify an SS to RS mapping file so that *graf* can convert the SS IDs to RS IDs.

Example: `graf -plink DsWithSs –SsToRs.txt (`assuming PLINK set DsWithSs.* exists)

OUTPUT FILES

graf requires that an input genotype file, either in PLINK format (with -plink option) or in GRAF format (with -geno option) should be specified. When -exfp option is selected, the -out option should also be selected to specify the name of the output file. The output file is the genotype dataset in GRAF format, as described above.

The output file should have the extracted genotypes of the fingerprinting SNPs and can be passed back to *graf* as an input file in a later run.

When -exfp option is not selected, *graf* will use the genotype information in the input genotype dataset, find the related samples and subjects, and will save the results to the output file.

If any related subjects are found, the results will be saved to the output file, which is a plain text file with the following columns:

Sample1: the ID of the first sample in each pair
Sample2: the ID of the second sample in each pair
Subject1: the subject ID of the first sample in each pair
Subject2: the subject ID of the second sample in each pair
Sex11: the gender of the first subject in each pair, 1=male; 2=female
Sex12: the gender of the second subject in each pair, 1=male; 2=female
HG match: number of SNPs with matched genotypes when only homozygous SNPs are counted
HG miss: number of SNPs with mismatched genotypes when only homozygous SNPs are counted
HGMR: Homozygous Genotype Mismatch Rate (%)
AG match: number of SNPs with matched genotypes when all SNPs are counted
AG miss: number of SNPs with mismatched genotypes when all SNPs are counted
AGMR: All Genotype Mismatch Rate (%)
Geno relation: relationship determined by sample genotypes. See above for code meanings
Ped relation: relationship derived from subject-sample mapping file and pedigree file. See above for code meanings
p_value: Probability that the genetic relationship is NOT the predicted type

When multiple PLINK sets are checked pairwise, the output file will have two extra columns, DS1 and DS2, showing the dataset IDs for the pair of PLINK sets.

REPORTING QUESTIONS AND PROBLEMS

Send any questions or problem reports to dbgap-help@ncbi.nlm.nih.gov.

There are two issues related to the input files for GRAF that can make filing problem reports for GRAF more complicated than for some other bioinformatics software packages: 1) the files may be large and 2) the files likely contain human subjects data that the user may not be authorized to send in the context of a problem report.

To avoid the two numbered problems, users should avoid sending input files with the initial problem report. Sending copies of diagnostic messages and/or screenshots of unexpected behavior is appropriate in the initial report and this is generally a good way for a user to report exactly what the user observed.

Some input files to GRAF may come from the dbGaP data repository. If a user wishes to ask a question concerning files from a dbGaP dataset, then the user can refer to the dbGaP accessions such as those starting with phs, phg, pht, the files names, and the versions, if any. Since dbgap-help personnel already have access to these files, there is no need or benefit to transmitting the dbGaP files back to dbgap-help.

For questions or problem reports that concern input files that are not from dbGaP (e.g., from EGA), users should follow a two-step procedure. Step 1: If the consents are such that the dbGaP staff has permission to look at the files, include in the initial problem report a request to transfer files via sftp. Step 2: The initial response from dbgap-help should include instructions on how to transfer the files, and the user should follow those instructions.

PLOTING GRAF RESULTS USING *PlotGraf.pl*

PlotGraf.pl is a perl script that plots graphs to show the distributions of HGMR and AGMR values. It shows brief instructions when it is executed without parameters:

```
PlotGraf.pl

Usage: PlotGraf.pl <input related subject file> <output png file>
                   <graph type> [Options]

Note:
    Valid graph types are:
        1 = HGMR histogram
        2 = AGMR histogram
        3 = HGMR + AGMR scatter plot

Options:
    -gw     graph width:  Set graph width in pixels
    -gh     graph height: Set graph height in pixels
    -xmax   max x value:  Set maximum HGMR or AGMR on x-axis of the
                          histogram
    -ymax   max y value:  Set maximum number of pairs on y-axis of the
                          histogram
    -dot    size:         Set dot size in pixels on the scatter plot
    -hfd    size:         Set dot size in pixels for HF (half sibling
                          + full cousin) pairs
```

It takes three required parameters. The first parameter should be the name of the file that is generated by *graf* and contains related subject pairs. The second one is the output .png file which shows the graph. The third one is an integer representing the graph type. The options should be entered after the required parameters. Below are some examples showing how to run the script.

```
graf -plink affy_hapmap -maxhm 15 -ssm affy_hapmap_ssm.txt
     -ped affy_hapmap_pedigree.txt -out affy_hapmap_rels.txt
PlotGraf.pl affy_hapmap_rels.txt affy_hapmap_hgmr.png 1
PlotGraf.pl affy_hapmap_rels.txt affy_hapmap_agmr.png 2
PlotGraf.pl affy_hapmap_rels.txt affy_hapmap_scatter.png 3
```

In the first step the C++ program finds related pairs and saves the results to affy_hapmap_rels.txt. Then PlotGraf.pl takes the results and plots histograms to show distributions of HGMR values of the related subjects, AGMR values of the duplicates, and a scatter plot to show distribution of both values.

In both histograms, the colored bars represent different type of relationships derived from the SSM and pedigree file (See Table 2 for the meanings of the 2-letter abbreviations). The cyan lines show the cutoff values suggested by GRAF to separate different types of relationships determined by comparing the genotypes. In the scatter plot, each contour line shows the area that is predicted to contain 95% of the pairs for each relatedness type, assuming all of the 10,000 fingerprinting SNPs are genotyped for all of the subjects in a large, homogeneous, random mating population. Note that the HapMap samples were collected from human individuals from very different populations, and GRAF is more accurate when predicting relatedness for subjects from a homogeneous population.

```
graf -plink affy_hapmap -maxhm 15 -ssm affy_hapmap_fake_ssm.txt
     -ped affy_hapmap_fake_pedigree.txt -out affy_hapmap_fake_rels.txt
PlotGraf.pl affy_hapmap_fake_rels.txt affy_hapmap_hgmr_f1.png 1
     -gw 1000 -gh 500
PlotGraf.pl affy_hapmap_fake_rels.txt affy_hapmap_agmr_f1.png 2
     -xmax 60 -ymax 20
PlotGraf.pl affy_hapmap_fake_rels.txt affy_hapmap_scatter_f1.png 3
     -dot 5
```

The above examples show that graph size, axis limits and the scatter plot dot size can be adjusted by users. In the first step a fake pedigree and a fake SSM file are used to show how GRAF finds and reports errors in the pedigree and SSM files. The HGMR histogram generated in the second step shows that some of the related pairs reported by the pedigree and SSM file don't match the genetic relatedness determined by GRAF. It also shows that the graph size can be adjusted by using options -gw and -gh. The AGMR histogram also shows the mismatches between the relationships types reported in the input files and those determined by GRAF. The axis limits can be adjusted by using -xmax and -ymax options. The scatter plot shows the dot size can be adjusted using -dot option.

Multiple genotype datasets can be combined into one .fpg file and passed to *graf* for determining genetic relationships, e.g.,

```
graf -exfp affy_hapmap,perlegen_hapmap -out comb_hapmap.fpg
graf -geno comb_hapmap.fpg -out comb_hapmap_rels.txt -maxhm 15
     -ped affy_hapmap_pedigree.txt -ssm comb_hapmap_ssm.txt
PlotGraf.pl comb_hapmap_rels.txt comb_hapmap_hgmr.png 1
PlotGraf.pl comb_hapmap_rels.txt comb_hapmap_agmr.png 2
PlotGraf.pl comb_hapmap_rels.txt comb_hapmap_scatter.png 3
```

When multiple datasets are used, if there are no SSM and pedigree files, it is not required that the sample and subject IDs be unique across datasets. GRAF uses both DS# and subject/sample IDs to identify subjects or samples. In the output table, GRAF shows both the DS# and ID for each subject or sample. However, when there are SSM and pedigree files, it is required that IDs be unique across datasets. GRAF doesn't take multiple SSM or pedigree files. The user needs to combine multiple SSM or pedigree files into one, and each ID in the combined SSM or pedigree file should represent only one sample or subject. Neither the SSM file nor the pedigree file has DS# columns.

The -hfd option of PlotGraf.pl lets user set the dot size for the half sibling + full cousin pairs (HF, see Table 2) in the scatter plot. The HF relationship is genetically remoter than full sibling but closer than second degree relatives. In the scatter plot, these pairs are predicted to be between FS and D2 pairs. In the rare cases when there are HF pairs, the user can use -hfd option to highlight the HF pairs by setting different dot sizes for them.

METHODS

Let A and B be the two alleles of each SNP, for each pair of samples, HGMR is calculated using the following equation:

$$\text{HGMR} = \frac{N_{AA/BB}+N_{BB/AA}}{N_{AA/BB}+N_{BB/AA}+N_{AA/AA}+ N_{BB/BB}}$$

where $N_{AA/BB}$, $N_{BB/AA}$, $N_{AA/AA}$, $N_{BB/BB}$ are the counts of SNPs with these genotypes.

Both AGMR and HGMR can be used to distinguish subject pairs with different relationships from each other. Suppose there is a homogeneous, random mating population in Hardy-Weinberg equilibrium, and the frequencies of the two alleles at a biallelic SNP location *i* are p_i and q_i. Assume *S* independent SNPs are compared, the expected AGMR and HGMR values of subject pairs with relationships identical twins (ID), parent-offspring (PO), full sibling (FS), second degree relative (D2), third degree relative (D3), and unrelated subjects (UN) can be calculated using the following equations.

Let

$$P_{Z0} = \frac{1}{S}\sum_{i=1}^{s}(4p_i^2q_i^2 + p_i^4 + q_i^4) \tag{1}$$

$$P_{Z1} = \frac{1}{S}\sum_{i=1}^{s}(p_i^3 + q_i^3 + p_i^2q_i + p_iq_i^2) \tag{2}$$

$$P_{Z2} = 1 \tag{3}$$

The expected AGMR values for different types of relationships will be:

$$E(AGMR_{ID}) = 1 - P_{Z2} = 0 \tag{4}$$

$$E(AGMR_{PO}) = 1 - P_{Z1} \tag{5}$$

$$E(AGMR_{FS}) = 1 - (\frac{1}{4} P_{Z0} + \frac{1}{2} P_{Z1} + \frac{1}{4} P_{Z2}) \tag{6}$$

$$E(AGMR_{D2}) = 1 - (\frac{1}{2} P_{Z0} + \frac{1}{2} P_{Z1}) \tag{7}$$

$$E(AGMR_{D3}) = 1 - (\frac{3}{4} P_{Z0} + \frac{1}{4} P_{Z1}) \tag{8}$$

$$E(AGMR_{UN}) = 1 - P_{Z0} \tag{9}$$

and the expected HGMR values are:

$$E(HGMR_{ID}) = 0 \tag{10}$$

$$E(HGMR_{PO}) = 0 \tag{11}$$

$$E(HGMR_{FS}) = \frac{\sum_{i=1}^{s} 2p_i^2 q_i^2}{\sum_{i=1}^{s}[(2p_i^2 q_i^2 + p_i^4 + q_i^4) + 2(p_i^3 + q_i^3) + (p_i^2 + q_i^2)]} \tag{12}$$

$$E(HGMR_{D2}) = \frac{\sum_{i=1}^{s} 2p_i^2 q_i^2}{\sum_{i=1}^{s}(2p_i^2 q_i^2 + p_i^4 + q_i^4 + p_i^3 + q_i^3)} \tag{13}$$

$$E(HGMR_{D3}) = \frac{3\sum_{i=1}^{s} 2p_i^2 q_i^2}{\sum_{i=1}^{s}[3(2p_i^2 q_i^2 + p_i^4 + q_i^4) + (p_i^3 + q_i^3)]} \tag{14}$$

$$E(HGMR_{UN}) = \frac{\sum_{i=1}^{s} 2p_i^2 q_i^2}{\sum_{i=1}^{s}(2p_i^2 q_i^2 + p_i^4 + q_i^4)} \tag{15}$$

The above equations show that AGMR can be used to distinguish identical twins from pairs with other relationships, and HGMR can be used to separate parent-offspring pairs from subject pairs with relationships other than identical twins. Other related pairs can also be separated from each other using AGMR and HGMR, especially when both values are used.

Since HGMR and AGMR values of each relationship type except for parent-offspring and identical twins follow normal distributions, the combination of both variables should follow a bivariate normal distribution. In general, if *x* and *y* are two variables following normal distributions $x \sim N(\mu_x, \sigma_x^2)$ and $y \sim N(\mu_y, \sigma_y^2)$, then we have the following probability density function for both *x* and *y*:

$$P(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left(\frac{-1}{2(1-\rho^2)}\left[\frac{(\mathrm{x}-\mu_\mathrm{x})^2}{\sigma_\mathrm{x}^2} - \frac{2\rho(\mathrm{x}-\mu_\mathrm{x})(\mathrm{y}-\mu_\mathrm{y})}{\sigma_\mathrm{x}\sigma_\mathrm{y}} + \frac{(\mathrm{y}-\mu_\mathrm{y})^2}{\sigma_y^2}\right]\right) \qquad (16)$$

where ρ is the correlation coefficient between x and y:

$$\rho = \frac{covariance(x,y)}{\sigma_x\sigma_y}$$

Setting HGMR = x and AGMR = y, we estimate the values of μ_x, σ_x, μ_y, σ_y, and ρ for each relationship type using the phenotype and genotype data of all the subjects in the current (04/02/2015) dbGaP database and then use these values to determine the relationship between each pair of subjects. The results are shown in Table 1.

Equation (16) can be used to calculate the probability densities for relationships FS, D2, D3, and UN. It cannot be used for PO and ID pairs, since the HGMR values of PO pairs do not follow normal distributions. The homozygous genotype mismatches of PO pairs are all caused by genotyping errors. If genotyping errors are random over all the SNPs, then HGMR values will follow a Poisson distribution. Unfortunately, the actual genotyping errors are not random and the HGMR values of PO pairs don't follow Poisson distribution. Actually the number of pairs roughly decreases exponentially when HGMR increases. We use the following probability density function to estimate the distribution of both HGMR and AGMR values of PO pairs:

$$P_{PO}(x,y) = \frac{k}{\sigma\sqrt{2\pi}}e^{-kx-\frac{(y-\mu)^2}{2\sigma^2}} \qquad (17)$$

where μ and σ are from Table 2, and k is set to 0.1 based on the actual HGMR values of the PO pairs in dbGaP database.

Table 1. Mean HGMR and AGMR values and correlation coefficients between HGMR and AGMR of all related subjects reported in the data files submitted to dbGaP

	Number	HGMR (%)		AGMR (%)		Correlation
Relationship	of pairs	Mean	SD	Mean	SD	coefficient
Self or monozygotic twins	21099	0.00	0.00	0.08	0.38	0.317
Parent-offspring	45184	0.04	0.07	39.48	1.47	0.104
Full sibling	28447	4.86	1.02	33.59	2.32	0.784
Second degree relative	18492	11.16	1.35	47.30	1.10	0.803
Third degree relative	10841	17.32	1.35	51.07	0.96	0.830

Using Equations (16) and (17), we can predict the distribution HGMR and AGMR values of subject pairs of each relationship type. Figure 1 shows the predicted contour lines of different relationship types for a typical dbGaP study. The area within each contour line is expected to contain 95% of the subject pairs of one relationship type. Note that PO, FS and D2 pairs are separated from each other very well, but there are some overlaps between D2 and D3 pairs.

GRAF calculates HGMR and AGMR for each pair of subjects determines that a pair of samples is identical (either from the same subject or from identical twins) if the AGMR value is less than the cutoff value 18%. For the pairs that are not identical, GRAF uses Equation (16) to compute the probability density for each relationship type FS, D2 and D3, and uses Equation (17) to compute the probability density for relationship PO. Parameters σ_x, σ_y, and ρ are from Table 1, while μ_x and μ_y are calculated using Equations (1) to (15). GRAF compares the probability densities of all the four relationship types and determines the relationship by finding the type with the maximum probability density. Let x, y be the HGMR and AGMR values for one pair of subjects, respectively. Denote xy as the joint occurrence of HGMR = x and AGMR = y. For each relationship $R \in$ {PO, FS, D2, D3}, GRAF calculates the probability that the relationship is R assuming all pairs are either PO, FS, D2 or D3 and the numbers of pairs of these types are all equal, using the following equation:

$$P(R|xy) \approx \frac{P(xy|R)}{P(xy|PO)+P(xy|FS)+P(xy|D2)+P(xy|D3)} \tag{18}$$

P(xy|R) for each relationship *R* can be calculated using Equations (1) or (2). GRAF determines the relationship to be the one that maximizes *P(R|xy)*. Since *P(R|xy)* is usually close to 1, GRAF defines p-value as the probability that one pair DOES NOT have relationship *R*, and calculates p-value as 1 - *P(R|xy)* and reports this value to the users.

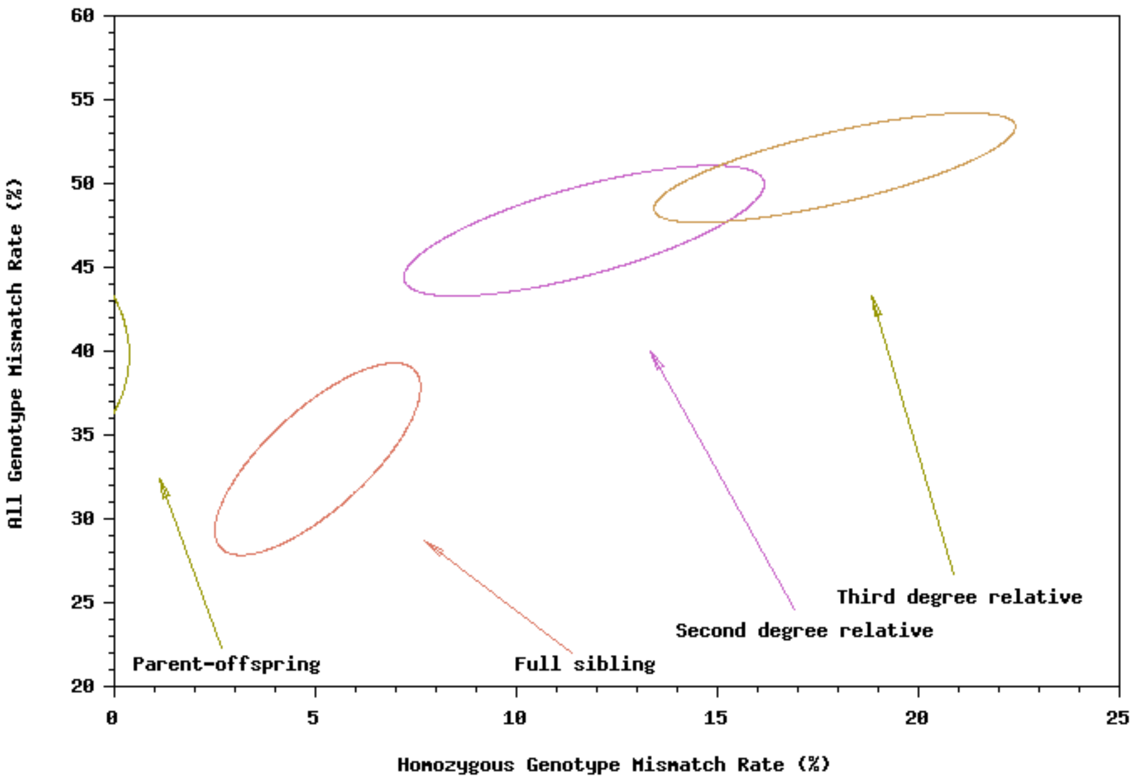


Figure 1. Expected Distribution of Homozygous Genotype Mismatch Rates and All Genotype Mismatch Rates. Each contour line shows the area predicted by Equations (16) and (17) as to contain 95% of the related pairs of one relationship type.

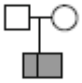
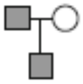
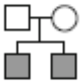
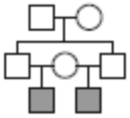
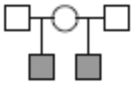
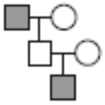
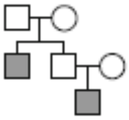
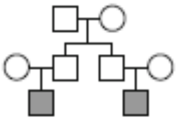
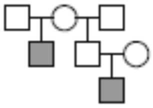
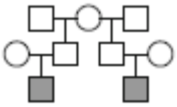
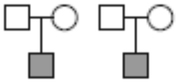
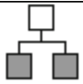
GRAF calculates HGMR and AGMR values for every pair of subjects and reports all the pairs that it considers to be third-degree relatives or closer. When a pedigree file is provided, GRAF checks the pedigree file to find all pairs of related subjects within three generations, validates them using genotype data and reports the subject pairs identified by GRAF as different types. Table 2 shows all types of relationships checked by GRAF and their expected relationships obtained by comparing their genotypes. GRAF reports a union of the following two sets:

1. All pairs determined by GRAF as related with 3rd degree or closer
2. All pairs derived from pedigree file with relationship 3rd degree or closer

For each pair of subjects in the union, GRAF shows both the relationship determined by the program and that reported in pedigree

Unexpected genetic relationships reported by GRAF may indicate errors in the pedigree file, SSM file or genotype datasets. However, it is also possible that some of the unexpected relationships are caused by a lack of genotyping information, or a lack of discrimination power of the program. GRAF reports those mismatches that are most likely caused by errors in the pedigree file or SSM file, or mislabeled sample IDs in the genotype datasets. The users are expected to check the results and make their own decisions.

Table 2. Pedigree relationships and the expected genetic relationships

Pedigree	Relationship	Abbr.	Expected genetic relation
	Self	DP	ID
	MZ-twin	MT	ID
	Parent-offspring	PO	PO
	Full sibling	FS	FS
	Half sibling + First cousin	HF	FS or D2
	Half sibling	HS	D2
	Grandparent-Grandchild	GP	D2
	Avuncular	AV	D2
	First cousin	FC	D3
	Half avuncular	HA	D3
	Half first cousin	HC	D4 (reported as UN)
	Unrelated	UN	UN
	Sibling with one parent ID missing	SB	FS or D2

REFERENCES

1. Jin, Y, Schäffer, A.A., Sherry, S.T., Feolo M. Quickly identifying identical and closely related subjects in large databases using genotype data. *PLoS One* **12(6):**e0179106 (2017).